Modeling Frequency and Severity of Claims with the Generalized Cluster-Weighted Model

Nikola Počuča, Tatjana Miljkovic, Petar Jevtić and Paul D. McNicholas

October 19, 2018

Abstract

In this paper, we propose a generalized cluster-weighted model (GCWM) that allows for modeling non-Gaussian distribution of the continuous covariates and a new zero-inflated GCWM (ZI-GCWM) for modeling insurance claims data with excess zeros. We describe two expectation-optimization (EM) algorithms for parameter estimation in GCWM and ZI-GCWM. A simulation study showed that both cluster models perform well for different settings in contrast to the existing mixture-based approaches. A real data set based on French automobile policies is used to illustrate the application of the proposed models.

KEY WORDS: GCWM, CWM, ZI-GCWM, clustering, automobile claims.

JEL CLASSIFICATION: C02, C40, C60.

1 Introduction

A significant number of clustering methods have been proposed for sub-grouping the data in the area of computer science, biology, social science, statistics, marketing, etc. Ingrassia et al. (2015) proposed a cluster-weighted model (CWM) framework as a flexible family of mixture models for fitting the joint distribution of a random vector composed of a response variable and a set of mixed-type covariates with the assumption that continuous covariates come from Gaussian distribution. CWMs with Gaussian assumptions have been proposed by Gershenfeld (1997), Gershenfeld et al. (1999), and Gershenfeld (1999) in a context of media technology. Some extensions of this class of models have been considered by Punzo and Ingrassia (2014), Ingrassia et al. (2014a), Ingrassia et al. (2014b), Subedi et al. (2013, 2015), and Punzo and McNicholas (2017). These

clustering methods are somewhat lacking for modelling insurance data, e.g., high excess zeros for claim count, heavy-tail loss distribution, deductible, or limits.

Sub-grouping of insurance policies based on risk classification is a standard practice in insurance. The heterogenous nature of insurance data allows for explorations of many different techniques for sub-grouping risk. As a result, there is a growing number of papers in the area of mixture modeling of univariate and multivariate insurance data to account for heterogeneity of risk. Lee and Lin (2010), Verbelen et al. (2015), and Miljkovic and Grün (2016) proposed mixture models for univariate loss data, and mixture modeling of univariate insurance data has been extended to the multivariate context. A finite mixture of bivariate Poisson regression models with an application to insurance ratemaking was studied by Bermúdez and Karlis (2012). A Poisson mixture model for count data was considered by Brown and Buckley (2015) with application in managing a Group Life insurance portfolio. Recently, Miljkovic and Fernández (2018) reviewed two complementary mixture-based clustering approaches (CWMs and mixture-based clustering for an ordered stereotype model) for modeling unobserved heterogeneity in an automobile insurance portfolio, depending on the data structure under consideration.

In this paper, we extend the CWM family proposed by Ingrassia et al. (2015) to allow for modeling of non-Gaussian continuous covariates and a zero-inflated Poisson (ZIP) claims data with excess zeros, which are commonly seen in the insurance applications. We consider a generalized cluster-weighted model (GCWM) as well as a zero-inflated GCWM (ZI-GCWM). Two partitioning methods are considered with two separate expectation-maximization (EM) algorithms (Dempster et al., 1977). The EM algorithm is based on the complete-data likelihood, which encompasses the observed data together with the missing data and/or latent variables. The EM algorithm can be highly effective for maximum likelihood estimation when data is incomplete or is assumed to be incomplete. The first EM algorithm is for parameter estimation for the GCWM models, while the second EM is for parameter estimation for the ZI-GCWM. We show that the Bernoulli and Poisson GCWM accurately estimate the initialization of the EM algorithm for the ZI-GCWM model. These models utilize individual claims data and should be useful in the areas of ratemaking and risk management.

This paper is organized as follows. The GCWM and GCWM approaches are discussed in Section 2, and parameter estimation is discussed in Section 3. Then, our methodology is applied to real data on French automobile claims and an extensive simulation study is conducted (Section 4). This paper concludes with a discussion and some suggestions for future work (Section 5).

2 Methodology

2.1 Background

Let $(\boldsymbol{X}', Y)'$ be the pair of a vector of covariates $\boldsymbol{X}$ and a response variable Y. Assume this pair is defined on some sample space Ω that takes values in an appropriate Euclidian subspace. Now, assume that there exists G non-overlapping partitions of Ω, denoted as $\Omega_1, \ldots, \Omega_G$. Gershenfeld (1997) characterized CWMs as a finite mixture of GLMs; hence, the joint distribution $(\boldsymbol{X}', Y)'$ has the form

$$f(\boldsymbol{x}, y; \boldsymbol{\Phi}) = \sum_{j=1}^{G} \tau_j q(y|\boldsymbol{x}; \boldsymbol{\vartheta}_j) p(\boldsymbol{x}; \boldsymbol{\vartheta}_j), \tag{2.1}$$

where $\boldsymbol{\Phi}$ denotes the model parameters. The pair $q(y|\boldsymbol{x}; \boldsymbol{\vartheta}_j)$ and $p(\boldsymbol{x}; \boldsymbol{\vartheta}_j)$ are conditional and marginal distributions of $(\boldsymbol{X}', Y)'$, respectively, while τ_j is the jth mixing proportion such that $\sum_{j=1}^{G} \tau_j = 1$, $\tau_j > 0$. Ingrassia et al. (2015) proposed a flexible family of mixture models for fitting the joint distribution of a random vector $(\boldsymbol{X}', Y)'$ by splitting the covariates into continuous and discrete, i.e., $\boldsymbol{X} = (\boldsymbol{V}', \boldsymbol{W}')'$. The assumption of independence between continuous and discrete covariates allows us to multiply their corresponding marginal distributions. Thus, for this setting the model in (2.1) is reformulated as follows

$$f(\boldsymbol{x}, y; \boldsymbol{\Phi}) = \sum_{j=1}^{G} \tau_j q(y|\boldsymbol{x}; \boldsymbol{\vartheta}_j) p(\boldsymbol{x}; \boldsymbol{\theta}_j) = \sum_{j=1}^{G} \tau_j q(y|\boldsymbol{x}; \boldsymbol{\vartheta}_j) p(\boldsymbol{v}; \boldsymbol{\theta}_j^{\star}) p(\boldsymbol{w}; \boldsymbol{\theta}_j^{\star\star}) \tag{2.2}$$

where $\boldsymbol{v}$ and $\boldsymbol{w}$ are the vectors of continuous and discrete covariates, respectively, $q(y|\boldsymbol{x}; \boldsymbol{\vartheta}_j)$ is the conditional density of $Y|\boldsymbol{x}$ with parameter vector $\boldsymbol{\vartheta}_j$, $p(\boldsymbol{v}; \boldsymbol{\theta}_j^{\star})$ is the marginal distribution of $\boldsymbol{v}$ with parameter vector $\boldsymbol{\theta}_j^{\star}$, and $p(\boldsymbol{w}; \boldsymbol{\theta}_j^{\star\star})$ is the marginal distribution of $\boldsymbol{w}$ with parameter vector $\boldsymbol{\theta}_j^{\star\star}$. As before, $\boldsymbol{\Phi}$ denotes the model parameters. Note that the conditional distribution $q(y|\boldsymbol{x}; \boldsymbol{\vartheta}_j)$ is assumed to belong to an exponential family of distributions and as such can be modeled in the GLM framework. Here, the marginal distribution of continuous covariates is assumed to be Gaussian. Unfortunately, this last assumption is too strong for use in insurance related applications, specifically in rate-making. To relax it, we develop the GCWM, which allows for non-Gaussian covariates as discussed in the next section.

2.2 Generalized cluster-weighted model (GCWM)

We proceed to extend (2.2) by splitting the continuous covariates $\boldsymbol{V}$ via $\boldsymbol{V} = (\boldsymbol{U}', \boldsymbol{T}')'$, where $\boldsymbol{U}$ contains the non-Gaussian covariates and $\boldsymbol{T}$ contains the Gaussian covariates. Thus (2.2) becomes

$$f(\boldsymbol{x}, y; \boldsymbol{\Phi}) = \sum_{j=1}^{G} \tau_j q(y|\boldsymbol{x}; \boldsymbol{\vartheta}_j) p(\boldsymbol{t}; \boldsymbol{\theta}_j^{\star}) p(\boldsymbol{w}; \boldsymbol{\theta}_j^{\star\star}) p(\boldsymbol{u}; \boldsymbol{\theta}_j^{\star\star\star}), \tag{2.3}$$

3

which we refer to as the GCWM. Here, $p(\boldsymbol{t}; \boldsymbol{\theta}_j^\star)$ denotes the marginal density of Gaussian covariates, with parameter vector $\boldsymbol{\theta}^\star$, and $p(\boldsymbol{u}; \boldsymbol{\theta}_j^{\star\star\star})$ is the marginal density of the non-Gaussian covariates with parameter vector $\boldsymbol{\theta}_j^{\star\star\star}$.

Because of its relevance to the actuarial application in this paper, we focus on the multivariate log-normal distribution for non-Gaussian covariates — this, however, does not reduce the generality of our general framework. With the log-normal assumption for $p(\boldsymbol{u}; \boldsymbol{\theta}_j^{\star\star\star})$, we have that $\boldsymbol{u}$ is defined on $\mathbb{R}_+^p$ with parameter vector $\boldsymbol{\theta}_j^{\star\star\star} = (\boldsymbol{\mu}_j^{\star\star\star}, \boldsymbol{\Sigma}_j^{\star\star\star})$ and probability density function

$$p\left(\boldsymbol{u}; \boldsymbol{\theta}_j^{\star\star\star}\right) = \frac{1}{(\prod_{i=1}^p u_i)|\boldsymbol{\Sigma}_j^{\star\star\star}|(2\pi)^{\frac{p}{2}}} \exp\left[-\frac{1}{2}(\ln \boldsymbol{u} - \boldsymbol{\mu}_j^{\star\star\star})' \boldsymbol{\Sigma}_j^{\star\star\star-1}(\ln \boldsymbol{u} - \boldsymbol{\mu}_j^{\star\star\star})\right]. \qquad (2.4)$$

The derivation of (2.4) can be found in the Appendix A.

2.3 Zero-inflated Poisson Model

In the zero-inflated Poisson (ZIP) model (see Lambert, 1992), we can split the conditional density of the response variable Y, i.e., $p(y|\boldsymbol{x}, \boldsymbol{\vartheta}_j)$, into zero and non-zero densities. The conditional probability mass associated with the event $y = 0$ is characterized by $q(y = 0|\boldsymbol{x}, \boldsymbol{\vartheta}_j)$. For situations when $y > 0$, the response variable Y is conditionally distributed with density $q(y > 0|\boldsymbol{x}, \boldsymbol{\vartheta}_j)$. Given the conditional density for the ZIP model, (2.3) can be re-written as

$$f(\boldsymbol{x}, y; \Phi) = \sum_{j=1}^G \tau_j \left[q(y = 0|\boldsymbol{x}; \boldsymbol{\vartheta}_j) + q(y > 0|\boldsymbol{x}; \boldsymbol{\vartheta}_j)\right] p(\boldsymbol{t}; \boldsymbol{\theta}_j^\star) p(\boldsymbol{w}; \boldsymbol{\theta}_j^{\star\star}) p(\boldsymbol{u}; \boldsymbol{\theta}_j^{\star\star\star}). \qquad (2.5)$$

Define $\tilde{\boldsymbol{x}} = [\boldsymbol{1}, \boldsymbol{x}]$, which contains the covariates together with a placeholder for the intercept in the GLM. Denote the Poisson conditional density as $q^P(y|\boldsymbol{x}; \lambda_j)$, where $y \in \{0, 1, \dots\}$, and let $\boldsymbol{\beta}_j$ be a row coefficient vector. The link function will be modelled with log-link for the GLM such that

$$\lambda_j = e^{\tilde{\boldsymbol{x}}\boldsymbol{\beta}_j'} \qquad \text{and} \qquad q^P(y|\boldsymbol{x}; \lambda_j) = e^{-\lambda_j} \frac{\lambda_j{}^y}{y!}.$$

Now, we use a Bernoulli model for the conditional density. We denote the density as $q^B(y|\boldsymbol{x}; \bar{\boldsymbol{\beta}}_j)$, where $\bar{\boldsymbol{\beta}}_j$ is a row coefficient vector. Here, the GLM will be modelled with the associated logit link function so that

$$\psi_j = \frac{e^{\tilde{\boldsymbol{x}}\bar{\boldsymbol{\beta}}_j'}}{1 + e^{\tilde{\boldsymbol{x}}\bar{\boldsymbol{\beta}}_j'}} \qquad \text{and} \qquad q^B(y|\boldsymbol{x}; \psi_j) = \begin{cases} \psi_j, & y = 0, \\ 1 - \psi_j, & y > 0. \end{cases}$$

Now, given a combination of two preceding models, we introduce the ZIP model in which zero counts come from two random variables. One is the Bernoulli random variable, which generates structural zeros, and the

other is the Poisson random variable. The coefficients $\boldsymbol{\vartheta}_j = \{\boldsymbol{\beta}_j, \bar{\boldsymbol{\beta}}_j\}$ correspond to the two above introduced conditional densities where the coefficients are estimated using a GLM as in Lambert (1992). The components of ZIP conditional density $q(y|\boldsymbol{x}; \boldsymbol{\vartheta}_j)$ are

$$q(y = 0|\boldsymbol{x}; \boldsymbol{\vartheta}_j) = \psi_j + (1 - \psi_j)e^{-\lambda_j} \qquad \text{and} \qquad q(y > 0|\boldsymbol{x}; \boldsymbol{\vartheta}_j) = (1 - \psi_j)e^{-\lambda_j}\frac{(\lambda_j)^y}{y!}.$$

Also, the link functions we consider are log-link for the Poisson and logit link for the Bernoulli model so that

$$\psi_j = \frac{e^{\tilde{\boldsymbol{x}}\bar{\boldsymbol{\beta}}'_j}}{1 + e^{\tilde{\boldsymbol{x}}\bar{\boldsymbol{\beta}}'_j}} \qquad \text{and} \qquad \lambda_j = e^{\tilde{\boldsymbol{x}}\boldsymbol{\beta}'_j}. \tag{2.6}$$

Noe that, here, the parameter ψ_j denotes the mean of the Bernoulli distribution of the jth component from which extra zeros emanate, and the parameter λ_j characterizes the jth Poisson distribution. This allows for a more nuanced approach to handling the inflation of zeros for automobile insurance (see Bermúdez and Karlis, 2012).

2.4 Bernoulli-Poisson Sample Space Partitioning

The single component ZIP model assumes that the inflated zeros emanate from both a Bernoulli and Poisson random variables while the non-zeros are assumed to come exclusively from the Poisson random variable. However, recent research extends the single component ZIP models to mixture models for heterogeneous count data with excess zeros (see Bermúdez and Karlis, 2012). In mixtures of ZIPs, zeros are assumed to come from multiple different Binomial and Poisson random variables. Difficulties are apparent during the maximization step of the EM algorithm when means of covariates are very close together (see Lim et al., 2014). However, misclassification error can be reduced using parsimonious models for the independent variables as in McNicholas and Murphy (2008). [Is this the paper we mean to cite?] [In your paper with TB Murphy found here: https://pdfs.semanticscholar.org/6e54/f621ff1c43c3bac0ba989953e7253b5fa4c0.pdf, now fixed] In this work, we propose a new method to rectify this problem and partition the dataset using Bernoulli and Poisson GCWMs. Furthermore, we construct a ZI-GCWM using the previously generated Bernoulli and Poisson GCWMs. In the first EM algorithm we estimate parameters pertaining to the GCWM under the assumption of a Poisson model and, separately, we carry out the same process under the assumption of a Bernoulli model. Using the obtained parameter estimates from the two separate applications of the EM algorithm, we set the initialization parameters for the second EM algorithm pertaining to parameter estimation of the ZI-GCWM. The work of Lambert (1992) specifies that the MLE estimates for the separate Poisson and Bernoulli models provide an excellent initial guess, allowing EM to converge quickly for ZIPs. The Bernoulli-Poisson sample space partitioning method consists of two separate EM algorithms. The first EM algorithm is for generating the GCWM models, while

the second EM is for optimizing the ZI-GCWM.

Recall, $(\boldsymbol{X}', Y)'$ to be a vector defined on some sample space Ω. By assumption, this sample space is partitioned into G non-overlapping sets such that their union constitutes it ie. $\Omega = \bigcup_{k=1}^{G} \Omega_k$. [I do not think this has been discussed. I am not sure it is helpful?] [In what precise sense does a set have a shape?]

For each set Ω_k, a Poisson or a zero-inflated Poisson model exists. In the case of a Poisson model for group k, it's sample space Ω_k^P, coincides with the sample space Ω_k such that $\Omega_k = \Omega_k^P$. For the case of a zero-inflated Poisson model for group k, let Ω_k^B be the sample space of the Bernoulli model on Ω_k. We decompose Ω_k^B into the space of excess zeros Ω_k^{BZ}, and the space of non-excess zeros Ω^{NZ} as $\Omega_k^B = \Omega_k^{BZ} \cup \Omega_k^{NZ}$. By definition non-excess zeros must emanate from a Poisson model thus Ω_k^{NZ} will coincide with the sample space of Ω_k^P such that $\Omega_k^{NZ} = \Omega_k^P$. Thus the $\Omega_k = \Omega_k^{BZ} \cup \Omega_k^P$, where Ω_k^{BZ} is the sample space of excess zeros emanating from a Bernoulli model, and Ω_k^P is the sample space relating to a Poisson model. Thus this construction allows both Poisson models, and zero-inflated Poisson models to exist on the same sample space Ω.

[I am not sure that all the material about sample spaces in the following is helpful here. Thoughts?] Specifically, if we introduce the Bernoulli model in a generalized form for conditional density (see Ingrassia et al., 2015), we have the sample space Ω^B and joint probability density function f^B to be

$$\Omega^B = \bigcup_{l=1}^{M \leq G} \Omega_l^B \qquad \text{and} \qquad f^B(\boldsymbol{x}, y; \Phi) = \sum_{l=1}^{M \leq G} \tau_l q^B(y|\boldsymbol{x}; \bar{\boldsymbol{\beta}}_l) p(\boldsymbol{t}; \boldsymbol{\theta}_l^\star) p(\boldsymbol{w}; \boldsymbol{\theta}_l^{\star\star}) p(\boldsymbol{u}; \boldsymbol{\theta}_l^{\star\star\star}).$$

Where the sample space Ω^B is partitioned up to $M \leq G$ non-overlapping sets. Similarly, if we introduce a Poisson model in a generalized form, the sample space Ω^P and joint probability density function f^P become

$$\Omega^P = \bigcup_{j=1}^{G} \Omega_j^P \qquad \text{and} \qquad f^P(\boldsymbol{x}, y; \Phi) = \sum_{j=1}^{G} \tau_j q^P(y|\boldsymbol{x}; \boldsymbol{\beta}_j) p(\boldsymbol{t}; \boldsymbol{\theta}_j^\star) p(\boldsymbol{w}; \boldsymbol{\theta}_j^{\star\star}) p(\boldsymbol{u}; \boldsymbol{\theta}_j^{\star\star\star}).$$

Where the sample space Ω^P is partitioned into G non-overlapping sets. Now, we can see a partitioning of a sample space Ω as

$$\Omega = \Omega^B \cup \Omega^P = \left(\bigcup_{l=1}^{M \leq G} \Omega_l^B \right) \bigcup \left(\bigcup_{j=1}^{G} \Omega_j^P \right) = \bigcup_{k=1}^{G} \left(\Omega_k^B \cup \Omega_k^P \right) := \bigcup_{k \in \{1,...,G\}} \Omega_k^Z,$$

there, the joint probability density function f^Z becomes

$$f^Z(\boldsymbol{x}, y, \Phi) = \sum_{k=1}^{G} \tau_k q_k^Z(y|\boldsymbol{x}; \bar{\boldsymbol{\beta}}_k, \boldsymbol{\beta}_k) p(\boldsymbol{t}; \boldsymbol{\theta}_k^\star) p(\boldsymbol{w}; \boldsymbol{\theta}_k^{\star\star}) p(\boldsymbol{u}; \boldsymbol{\theta}_k^{\star\star\star}).$$

Therefore the new conditional density for Ω_k^Z is now result of a model in which each component is captured by the conditional probability density function that is a mixture of particular Bernoulli and particular Poisson

6

densities

$$q_k^Z(y|\boldsymbol{x};\bar{\boldsymbol{\beta}}_k,\boldsymbol{\beta}_k) := q^B(y|\boldsymbol{x};\bar{\boldsymbol{\beta}}_k) + (1 - q^B(y|\boldsymbol{x};\bar{\boldsymbol{\beta}}_k))q^P(y|\boldsymbol{x};\boldsymbol{\beta}_k)$$

$$= q(y = 0|\boldsymbol{x};\boldsymbol{\vartheta}_k) + q(y > 0|\boldsymbol{x};\boldsymbol{\vartheta}_k), \quad k \in \{1, ..., G\}. \tag{2.7}$$

The initialization parameters for the second EM algorithm are provided by Bernoulli and Poisson GCWMs from (2.7) giving parameter pairs (ψ_k, λ_k). The second EM procedure then optimizes the ZI-GCWM. The ZI-GCWM is compared against the standard Poisson GCWM using a likelihood ratio test which is discussed in Section 3.3.

3 Parameter Estimation

The common approach for estimating parameters in finite mixture models is based on the EM algorithm (see McNicholas, 2016, for examples). The estimation of the developed Bernoulli-Poisson partitioning method is split into two EM algorithms. The first EM algorithm partitions the sample space, while the second EM algorithm optimizes the zero inflated portion.

3.1 EM Algorithm for Partitioning of Sample Space

The EM algorithm is based on the local maximum likelihood estimation. The initial values of the parameter estimates can be generated from a variety of strategies outlined in Biernacki et al. (2000). The algorithm proceeds by alternation of the E- and M-steps to update parameter estimates. The convergence criterion of the EM algorithm is based on the Aitken acceleration. It is used to estimate the asymptotic maximum of the log-likelihood at each iteration of the EM algorithm when the relative increase in the log-likelihood function is no bigger than a small pre-specified tolerance value or the number of iterations reach a limit. To find an optimal number of components, maximum likelihood estimation is obtained over a range of G groups, and the best model is selected based on the Bayesian information criterion (BIC).

In this subsection, we explain the parameter estimation in line with the GCWM methodology proposed by Ingrassia et al. (2015). The proposed GCWM is based on the assumption that $q(y|\boldsymbol{x}, \boldsymbol{\vartheta}_j)$ belongs to the exponential family of distributions that are strictly related to GLMs. The link function defines the relationship between the linear predictor and the expected value of the distribution function as $g(\boldsymbol{\mu}_j) = \tilde{\boldsymbol{x}}\boldsymbol{\beta}'$, where $g(\boldsymbol{\mu}_j)$ is the link function. We are interested in the estimation of the vector $\boldsymbol{\beta}_j$, thus the distribution of $Y|\boldsymbol{x}$ is denoted by $q(y|\boldsymbol{x};\boldsymbol{\beta}_j, \lambda_j)$, where λ_j signifies an additional parameter to account for when a distribution belongs to a two-parameter exponential family.

7

The marginal distribution $p(\boldsymbol{x}; \boldsymbol{\theta}_j)$ has the following components: $p(\boldsymbol{t}; \boldsymbol{\theta}_j^\star)$, $p(\boldsymbol{w}; \boldsymbol{\theta}_j^{\star\star})$, and $p(\boldsymbol{u}; \boldsymbol{\theta}_j^{\star\star\star})$. The first marginal density $p(\boldsymbol{t}; \boldsymbol{\theta}_j^\star := (\boldsymbol{\mu}_j^\star, \boldsymbol{\Sigma}_j^\star))$ is modeled as a Gaussian distribution with mean $\boldsymbol{\mu}_j^\star$ and covariance matrix $\boldsymbol{\Sigma}_j^\star$. The marginal density of discrete covaraites $p(\boldsymbol{w}; \boldsymbol{\theta}_j^{\star\star})$ is assumed to have for each finite discrete covariate in $\boldsymbol{W}$, a representative binary vector $\boldsymbol{w}^r = (w^{r1}, \ldots, w^{rc_r})'$, where $w^{rs} = 1$ if $w_r = s \in \{1, \ldots, c_r\}$, and $w^{rs} = 0$ otherwise. [As mentioned before, the following equation seems to come out of nowhere.] Given the preceding assumptions about discrete covariates, the marginal density is written as

$$p(\boldsymbol{w}; \boldsymbol{\gamma_j}) = \prod_{r=1}^{d} \prod_{s=1}^{c_r} (\gamma_{jrs})^{w^{rs}} \tag{3.1}$$

for $j = 1, \ldots, G$, where $\boldsymbol{\gamma}_j = (\boldsymbol{\gamma}_{j1}', \ldots, \boldsymbol{\gamma}_{jd}')'$, $\boldsymbol{\gamma}_{jr} = (\boldsymbol{\gamma}_{jr1}', \ldots, \boldsymbol{\gamma}_{jrc_d}')'$, $\gamma_{jrs} > 0$, and $\sum_{s=1}^{c_r} \gamma_{jrs}$, $r = 1, \ldots, q$. The density $p(\boldsymbol{w}, \boldsymbol{\gamma}_j)$ represents the product of d conditionally independent multinomial distributions with parameters $\boldsymbol{\gamma}_{jr}$, $r = 1, \ldots, d$. Finally, the third marginal density $p(\boldsymbol{u}; \boldsymbol{\theta}_j^{\star\star\star})$ will be modelled with a multivariate log-normal distribution having a location parameter vector $\boldsymbol{\mu}_j^{\star\star\star}$ and scale parameter matrix $\boldsymbol{\Sigma}_j^{\star\star\star}$.

Let $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)$ be a sample of n independent observations drawn from model in (2.3). Consider a latent random variable Z_{ij}. The realization z_{ij} of the latent indicator variable takes the value of $z_{ij} = 1$ indicating that observation $(\boldsymbol{x_i}, y_i)$ originated from the jth mixture component and $z_{ij} = 0$ otherwise. [Need to explain complete-data earlier.] [I now understand, you meant introduce it in writing, not in definition]

Given the sample, the complete-data likelihood function $L_c(\boldsymbol{\Phi})$ is given by

$$L_c(\boldsymbol{\Phi}) = \prod_{i=1}^{n} \prod_{j=1}^{G} \left[\tau_j q(y_i | x_i; \boldsymbol{\beta}_j, \lambda_j) p(t_i; \boldsymbol{\mu}_j^\star, \boldsymbol{\Sigma}_j^\star) p(w_i; \gamma_j) p(u_i; \boldsymbol{\mu}_j^{\star\star\star}, \boldsymbol{\Sigma}_j^{\star\star\star}) \right]^{z_{ij}}, \tag{3.2}$$

Taking the logarithm of (3.2), the complete-data log-likelihood is

$$\ell_c(\boldsymbol{\Phi}) = \sum_{i=1}^{n} \sum_{j=1}^{G} z_{ij} \Big[\log(\tau_j) + \log q(y_i | x_i; \boldsymbol{\beta}_j, \lambda_j) + \tag{3.3}$$
$$\log p(t_i; \boldsymbol{\mu}_j^\star, \boldsymbol{\Sigma}_j^\star) + \log p(w_i; \boldsymbol{\gamma}_j) + \log p(u_i; \boldsymbol{\mu}_j^{\star\star\star}, \boldsymbol{\Sigma}_j^{\star\star\star}) \Big].$$

On the $(s+1)$th iteration, the E-step requires calculation of the conditional expectation of $\ell_c(\boldsymbol{\Phi})$. Because $\ell_c(\boldsymbol{\Phi})$ is linear with respect to z_{ij}, we simplify the calculation to the current expectation of Z_{ij}, where Z_{ij} is the random variable corresponding to the realization z_{ij}. Given the previous parameters $\boldsymbol{\Phi}^{(s)}$ and the observed data, we calculate the current conditional expectation of Z_{ij} as

$$\pi_{ij}^{(s)} = E[Z_{ij} | (\boldsymbol{x_i}, y_i); \boldsymbol{\Phi}^{(s)}]$$
$$= \frac{\tau_j^{(s)} q(y_i | x_i; \boldsymbol{\beta}_j^{(s)}, \lambda_j^{(s)}) p(t_i; \boldsymbol{\mu}_j^{\star(s)}, \boldsymbol{\Sigma}_j^{\star(s)}) p(w_i; \boldsymbol{\gamma}_j^{(s)}) p(u_i; \boldsymbol{\mu}_j^{\star\star\star(s)}, \boldsymbol{\Sigma}_j^{\star\star\star(s)})}{f(\boldsymbol{x}_i, y_i; \boldsymbol{\Phi}^{(s)})}.$$

On the M-step of the $(s+1)$th iteration, the conditional expectation of $\ell_c(\boldsymbol{\Phi})$ denoted as a function $Q(\boldsymbol{\Phi}|\boldsymbol{\Phi}^{(s)})$ is maximized with respect to $\boldsymbol{\Phi}$ where [What does the previous sentence mean? It seems to me that $Q(\boldsymbol{\Phi}|\boldsymbol{\Phi}^{(s)})$ is simply what is calculated on the E-step; however, what is written here suggests something quite different.] the values of z_{ij} in (3.3) are replaced by their current expectations π_{ij} yielding

$$
\begin{aligned}
Q(\boldsymbol{\Phi}|\boldsymbol{\Phi}^{(s)}) &= \sum_{i=1}^{n}\sum_{j=1}^{G}\pi_{ij}^{(s)}\big[\log(\tau_j) + \log q(y_i|x_i;\boldsymbol{\beta}_j,\lambda_j) + \log p(t_i;\boldsymbol{\mu}_j^\star,\boldsymbol{\Sigma}_j^\star) + \log p(w_i;\boldsymbol{\gamma}_j) \\
&\quad + \log p(u_i;\boldsymbol{\mu}_j^{\star\star\star},\boldsymbol{\Sigma}_j^{\star\star\star})\big] \\
&= \sum_{i=1}^{n}\sum_{j=1}^{G}\pi_{ij}^{(s)}\log(\tau_j) + \sum_{i=1}^{n}\sum_{j=1}^{G}\pi_{ij}^{(s)}\log q(y_i|x_i;\boldsymbol{\beta}_j,\lambda_j) + \sum_{i=1}^{n}\sum_{j=1}^{G}\pi_{ij}^{(s)}\log p(t_i;\boldsymbol{\mu}_j^\star,\boldsymbol{\Sigma}_j) \\
&\quad + \sum_{i=1}^{n}\sum_{j=1}^{G}\pi_{ij}^{(s)}\log p(w_i;\boldsymbol{\gamma}_j) + \sum_{i=1}^{n}\sum_{j=1}^{G}\pi_{ij}^{(s)}\log p(u_i;\boldsymbol{\mu}_j^{\star\star\star},\boldsymbol{\Sigma}_j^{\star\star\star}).
\end{aligned}
\tag{3.4}
$$

The M-step requires maximization of the Q-function with respect to $\boldsymbol{\Phi}$ which can be done separately for each term on the right hand side in (3.4). As a result, the parameter updates on the $(s+1)$th iteration are

$$
\hat{\tau}_j^{(s+1)} = \frac{1}{n}\sum_{i=1}^{n}\pi_{ij}^{(s)}, \qquad \hat{\boldsymbol{\mu}}_j^{\star(s+1)} = \frac{1}{\sum_{i=1}^{n}\pi_{ij}^{(s)}}\sum_{i=1}^{n}\pi_{ij}^{(s)}\boldsymbol{t}_i, \qquad \hat{\gamma}_{jr}^{(s+1)} = \frac{\sum_{i=1}^{n}\pi_{ij}^{(s)}\omega_i^{rs}}{\sum_{i=1}^{n}\pi_{ij}^{(s)}},
$$

$$
\widehat{\boldsymbol{\Sigma}}_j^{\star(s+1)} = \frac{1}{\sum_{i=1}^{n}\pi_{ij}^{(s)}}\sum_{i=1}^{n}\pi_{ij}^{(s)}(\boldsymbol{t}_i - \hat{\boldsymbol{\mu}}_j^{(s+1)})(\boldsymbol{t}_i - \hat{\boldsymbol{\mu}}_j^{(s+1)})'.
$$

Parameter updates for the log-normal distribution are as follows

$$
\hat{\boldsymbol{\mu}}_j^{\star\star\star(s+1)} = \frac{1}{\sum_{i=1}^{n}\pi_{ij}^{(s)}}\sum_{i=1}^{n}\pi_{ij}^{(s)}\ln \boldsymbol{u}_i,
$$

$$
\widehat{\boldsymbol{\Sigma}}_j^{\star\star\star(s+1)} = \frac{1}{\sum_{i=1}^{n}\pi_{ij}^{(s)}}\sum_{i=1}^{n}\pi_{ij}^{(s)}(\ln \boldsymbol{u}_i - \hat{\boldsymbol{\mu}}_j^{\star\star\star(s+1)})(\ln \boldsymbol{u}_i - \hat{\boldsymbol{\mu}}_j^{\star\star\star(s+1)})'.
$$

The updates for $\boldsymbol{\beta}_j$ are computed by maximizing the G conditional density terms of the form [Note sure what the previous sen. means?]

$$
\sum_{i=1}^{n}\pi_{ij}^{(s)}\log q(y_i|\boldsymbol{x}_i;\boldsymbol{\beta}_j,\lambda_j).
\tag{3.5}
$$

The numerical optimization for each term is discussed in Wedel and De Sabro (1995) and Wedel (2002). For additional implementation information, the reader is referred to the manual of the `flexCWM` package (Ingrassia et al., 2015) for R (R Core Team, 2018).

3.2 EM Algorithm for Zero-inflated Model

For a zero-inflated model, the EM-algorithm follows a similar procedure as above to optimize the conditional density given in (2.7). [But what is its purpose? What new information, if any, is the reader supposed to get here? What would be lost if it were just deleted?] [What is the purpose of the previous sentence? What does it add?] Specifically, the log-likelihood function of ψ_k and λ_k is expressed as

$$l(\psi_k, \lambda_k | y, \boldsymbol{x}) = \sum_{y_i=0} \log\left[e^{\tilde{\boldsymbol{x}}_i \bar{\boldsymbol{\beta}}_k'} + \exp\left(-e^{-\tilde{\boldsymbol{x}}_i \boldsymbol{\beta}_k'}\right)\right] + \sum_{y_i>0} \left(y_i \tilde{\boldsymbol{x}}_i \boldsymbol{\beta}_k' + e^{\tilde{\boldsymbol{x}}_i \boldsymbol{\beta}_k'}\right)$$
$$- \sum_{i=1}^n \log\left(1 + e^{\tilde{\boldsymbol{x}}_i \bar{\boldsymbol{\beta}}_k'}\right) - \sum_{y_i>0} \log(y_i!),$$

where y_i, and $\tilde{\boldsymbol{x}}_i$ refers to the ith row of the response variable y and covariate vector $\tilde{\boldsymbol{x}}$. [I do not see any mention of "covariate matrix" in Section 2.3 or anywhere else in the paper before this point. I do not see how we have a matrix of covariates?] Due to the first term, the log-likelihood function is rather complicated to maximize, however Lambert (1992) gives a meaningful solution. Consider a random variable $Z_{ik}^\star$ indicating with $z_{ik}^\star = 1$ when y_i is generated from the Bernoulli random variable of partition k, and $z_{ik}^\star = 0$ when y_i is generated from the Poisson random variable of the same partition. [Please use a letter other than O, perhaps W? Random variables are usually taken from near the end of the alphabet, i.e., $U, V, \ldots, Z$, and constants from near the start. Also, note that $\mathcal{O}$ is usually used for "order", as in big-O notation.] [What is $\mathcal{O}_{ik}$? The notation here suggests it is not a random variable.] Then, the complete-data log-likelihood is

$$l_c(\psi_k, \lambda_k | y, \boldsymbol{x}, \boldsymbol{z}_k^\star) = \sum_{i=1}^n \left(z_{ik}^\star \tilde{\boldsymbol{x}}_i \bar{\boldsymbol{\beta}}_k^{'} - \log\left(1 + e^{\tilde{\boldsymbol{x}}_i \bar{\boldsymbol{\beta}}_k^{'}}\right)\right)$$
$$+ \sum_{i=1}^n (1 - z_{ik}^\star)(y_i \tilde{\boldsymbol{x}}_i \boldsymbol{\beta}_k^{'} - e^{\tilde{\boldsymbol{x}}_i \boldsymbol{\beta}_k^{'}}) + \sum_{i=1}^n (1 - z_{ik}^\star) \log(y_i!)$$
$$= l_c(\psi_k; y, \boldsymbol{x}, \boldsymbol{z}_k^\star) + l_c(\lambda_k; y, \boldsymbol{x}, \boldsymbol{z}_k^\star) + \sum_{i=1}^n (1 - z_{ik}^\star) \log(y_i!),$$

where $\boldsymbol{z}_k^\star := [z_{1k}^\star, ..., z_{nk}^\star]$ is a realization of $\boldsymbol{Z}_k^\star := [Z_{1k}^\star, ..., Z_{nk}^\star]$.

Note that $l_c(\psi_k, \lambda_k | y, \boldsymbol{x}, \boldsymbol{z}_k^\star)$ separates allowing the maximization of [Easier than what?] $l_c(\psi_k; y, \boldsymbol{x}, \boldsymbol{z}_k^\star)$ and $l_c(\lambda_k; y, \boldsymbol{x}, \boldsymbol{z}_k^\star)$ separately for parameters ψ_k and λ_k. With the EM algorithm, maximization of parameters are done iteratively between estimating $Z_{ik}^\star$ with its expectation under current estimates for λ_k and ψ_k (E-Step), and then maximizing the conditional expectation of the complete-data log-likelihood (M-Step). [I don't know what this last sentence means.]

In the E-step, using current estimates $\psi_k^{(s)}$ and $\lambda_k^{(s)}$ from the partition Ω_k^Z, we calculate the expected value

of $\boldsymbol{Z}_{ik}^{\star}$ by its posterior mean $\hat{z^{\star}}_{ik}^{(s)}$ for each cluster k at iteration s as

$$
\hat{z^{\star}}_{ik}^{(s)} = \begin{cases} \left[1 + \exp\left(-\tilde{\boldsymbol{x}}_i \bar{\boldsymbol{\beta}}_k^{'(s)} - e^{\tilde{\boldsymbol{x}}_i \boldsymbol{\beta}_k^{'(s)}} \right) \right]^{-1}, & y_i = 0 \\ 0 \quad, & y_i > 0. \end{cases}
$$

The M-Step can be split into the maximization of two complete data log-likelihoods and the $\hat{\boldsymbol{z}}^{\star}_k$ calculated from the previous iteration (s) as

$$
l_c(\lambda_k; y, \boldsymbol{x} | \hat{\boldsymbol{z}^{\star}}_k^{(s)}) = \sum_{i=1}^{n} (1 - \hat{z^{\star}}_{ik}^{(s)}) (y_i \tilde{\boldsymbol{x}}_i \boldsymbol{\beta}_k^{'} - e^{\tilde{\boldsymbol{x}}_i \boldsymbol{\beta}_k^{'}}). \tag{3.6}
$$

$$
l_c(\psi_k; y, \boldsymbol{x} | \hat{\boldsymbol{z}^{\star}}_k^{(s)}) = \sum_{i=1}^{n} \left(\hat{z^{\star}}_{ik}^{(s)} \tilde{\boldsymbol{x}}_i \bar{\boldsymbol{\beta}}_k^{'} - \log\left(1 + e^{\tilde{\boldsymbol{x}}_i \bar{\boldsymbol{\beta}}_k^{'}} \right) \right). \tag{3.7}
$$

[In the E-step, the expected value of the complete-data log-likelihood is computed. This leaves me wondering about why the M-step is being framed in terms of the complete-data log-likelihood.] The maximization of (3.6) for GLM coefficients λ_k can be carried out using a weighted log-linear Poisson regression with weights $1 - \hat{z^{\star}}_{ik}^{(s)}$ (see McCullagh and Nelder, 1989)), yielding $\lambda_k^{(s+1)}$. While the parameter ψ_k for (3.7) can be maximized over a gradient yielding $\psi_k^{(s+1)}$ (see Lambert, 1992). [What is "the parameter"?]

3.3 Comparing zero-inflated Models

Until recently, the usual test for comparing zero-inflated to non-zero inflated models has been the Vuong Test for non-nested models (Vuong, 1989). However, recent work by (see Wilson, 2015) has shown the misuse of this test for zero inflation where it is pointed out that the ZIP model is falsely defined as a non-nested model. Furthermore, the Vuong Test fails to identify evidence of zero-deflation leading to inconsistencies in the hypothesis test (see Wilson, 2015). [We need more detail about the "misuse".] To rectify this, Wilson and Einbeck (2018) show that it is sufficient to test for zero-modificiation in the form of a likelihood ratio test, where the hypotheses are

$$
H_0 : \psi_k = 0 \quad \text{vs.} \quad H_1 : \psi_k \neq 0,
$$

and the test statistic φ is given by

$$
\varphi = -2 \left[l(\tilde{\lambda}_k; y, \boldsymbol{x}) - l(\lambda_k, \psi_k; y, \boldsymbol{x}) \right]. \tag{3.8}
$$

I presume there are words missing here? Perhaps: The test statistic (3.8) The test statistic (3.8) is shown to follow a chi-squared distribution with m degrees of freedom (χ_m^2) and confidence level $\alpha = 0.10$, where m

is the number of covariates selected for the Bernoulli model in (2.6).[1]. [When we write "see" a reference, it should be clear what one will find there. In some cases, including the preceding sentence, it is not clear to me.] The function $l(\tilde{\lambda}_k; y, \boldsymbol{x})$ is the log-likelihood of a single component GCWM Poisson model on Ω_{Z_k} parameterized by $\tilde{\lambda}_k$. Recall that ψ_k is the zero-inflation paramater of the kth parition. [What is the purpose of this qualifier?] In our approach, we will be using (3.8) to test for evidence of zero-inflation on partition Ω_k, and then using BIC for model comparisons on Ω_k. This approach quickly determines if there is zero-inflation on partition Ω_k. When evidence of zero-inflation is established, we search for the best possible linear model using BIC.

4 Numerical Application

4.1 Dataset

The proposed methodology is illustrated on the French motor severity and frequency datasets by policy. These datasets are available as part of the R package CASdatasets (Dutang and Charpentier, 2016) and they were previously used by Charpentier (2014) who demonstrated various GLM modeling approaches for fitting frequency and severity. The French automobile portfolio consists of 413,169 motor third-party liability policies with the associated risk characteristics. The loss amounts by policy ID are also provided. In the following section all monetary units are considered to be of euro denomination as the dataset originates from France. [What is "the following text"? Table 1 is not referenced in the text. Each table and each figure must be referenced within the text.]

4.2 Discussion and Results

4.2.1 Modelling Severity

In this section, we show improved results for GCWM over CWM in modeling French motor losses. Furthermore we investigate the results of the GCWM model for the valuation of heterogeneous risk. We consider the following covariates: population density ($Density$), driver age ($DriverAge$), car age ($CarAge$), car power level ($Power$), gas (Gas), and geographical region in France ($Region$). [I do not like this way of expressing the model. It seems very informal. I think covariates written like X_{Region}, with β_j coefficients, would be better.] [You already said this in math, don't repeat.] The $CarAge$ is modelled as a categorical variable with five categories: $[0, 1)$, $[1, 5)$, $[5, 10)$, $[10, 15)$, and 15+. Additionally, $DriverAge$ is modelled as a categorical variable

---

[1]See Liklihood Ratio Tests in Wilson and Einbeck (2018)

Table 1: The description of variables in the French Motor Third-Part Liability dataset.

Attribute	Description
Policy ID	Unique identifier of the policy holder
Claim Nb	Number of claims during exposure period (0,1,2,3,4)
Exposure	The exposure of policy in years (0–1.5)
Power	Power level of car ordered categorical (12 levels)
Car Age	Car age in years
Driver Age	Age of a legal driver
Brand	Car brands (7 types)
Gas	Diesel or Regular
Region	Regions in France (10 classifications)
Density	Number of inhabitants per km^2
Loss Amount	Portion of claim the insurance policy pays

with five categories: $[18, 23)$, $[23, 27)$, $[27, 43)$, $[43, 75)$, and $75+$. $Power$ is modelled into three categories as in Charpentier (2014): DEF, GH, and other. The fitted model is defined with the following expression

$$g(\mathbb{E}\left[Y_{LossAmount}|\boldsymbol{x}\right]) = \beta_{j0} + \beta_{jDensity}x_{Density} + \beta_{jCarAge}x_{CarAge} + \beta_{jDriverAge}x_{DriverAge} +$$

$$\beta_{jRegion}x_{Region} + \beta_{jPower}x_{Power} + \beta_{jGas}x_{Gas} =: \tilde{\boldsymbol{x}}\boldsymbol{\beta}'_j. \tag{4.1}$$

The canonnical log-link function g is used for the GCWM in (4.1).

Beginning with the continuous covariate $Density$, we want to inspect the shape of its univariate data to see if it follows Gaussian distribution. The left-hand side of Figure 1 clearly revels that the the $Density$ is rather skewed right with several observations that report high value of density. This indicates a need for a transformation. The log-normal transformation clearly improves the fit (see the right side of Figure 1).

Figure 1: Density variable: Left figure shows the fit when Gaussian distribution is imposed (CMW approach) to highly skewed data. Right figure shows the fit when log-normal assumption is applied (GCWM approach).

Given the log-normal assumption, the result of the transformation is a better AIC and BIC for GCWM over CWM. [Need a fuller sentence here.] Table 2 shows a considerable difference in BIC and AIC comparing CWM and GCWM. The five component CWM with a BIC of 285,268 is significantly higher than the four component GCWM with a considerably lower BIC of 89,568.

We now investigate the results of GCWM in relation to the valuation of risk. For practical uses, finding clusters allows us to create different classifications of risk for various groups of drivers. [The previous sentence seems to me to contain the key point. I think this should be highlighted in the introduction and, insofar as possi-

Table 2: Comparison of AIC and BIC for CWM versus GCWM for G number of clusters estimated.
[Column 2 here is "k". What is "k"?]

Model	G	AIC	BIC
CWM	1	$352,470$	$352,661$
	2	$314,560$	$314,949$
	3	$301,223$	$301,812$
	4	$287,020$	$287,808$
	5	**$284,283$**	**$285,268$**
GCWM	1	$111,129$	$111,320$
	2	$90,039$	$90,428$
	3	$89,476$	$90,065$
	4	**$88,781$**	**$89,568$**
	5	$88,731$	$89,717$

ble, expanded upon here and maybe in the conclusions.] The results of the GCWM allows for the classification of drivers into various groups allowing one to assign different rates. [What following GCWM?]

After fitting the model, we then inspect the size of each cluster. The GCWM approach has chosen four components as the best model to represent the data. The size of each cluster is displayed in Table 3. Attention is brought to largest quantity of drivers that are grouped into Cluster 4. This accounts for 46% of all drivers and is fairly concentrated in the center of Figure 2. From the results we can create an insurance model with the distinct characteristics.

Table 3: Size and colours of clusters for the GCWM a model.

Cluster 1	Cluster 2	Cluster 3	Cluster 4
1,683	5,766	848	7,093
Red	Green	Blue	Teal

Figure 2: Showing clusters by color scheme: Cluster 1 - Red, Cluster 2 - Green, Cluster 3 - Blue, Cluster 4 - Teal for $Loss Amount$ vs $Density$ on a log scale. [The resolution of this figure could be higher. Similar for Figure 3.]

The Cluster 3 drivers have both low variability and low average cost in claims, thus can be insured with a lower rate than other drivers. From a risk management perspective this is the most ideal case as these claims have very low variance and cost. Cluster 4 drivers have also low variability but a higher average cost, thus they would have a rate higher than Cluster 3. Cluster 2 drivers have the next highest cost and variability of the clusters, these drivers are colored in green in Figure 1. The final cluster colored in red has the highest cost and variaiblity in claims out of any of the other clusters. From a risk management perspective this cluster would

14

have the highest rate.

Table 4 shows a breakdown of the types of drivers, ordered by volatility in descending order. Beginning with V1 volatility level, these drivers tend to have claims between €51 to €154, with a standard deviation of €13, and a mean of €79. That means that these drivers rarely exceed costs and tend to have very low volatility of claims. Moving onto V2 , these drivers have the second lowest level of volatility. Drivers in this range tend to have claims anywhere between €1, 039 to €1, 324, with a standard deviation of €52, and a mean of €1, 109. Proceeding to V3, its volatility in claims is greater than the preceding levels. Drivers in this cluster have claims anywhere between €221 to €8, 841, with a mean of €1, 687, and a standard deviation of €1, 284. Finally, V4 denotes the level of highest volatility. Claims in this level reach the highest recorded claim of €2, 036, 833, a mean of €9, 717, and a standard deviation of €64, 835. I think we should try to add monetary units; as a reviewer, this would annoy me.]

I think we should try to add monetary units; as a reviewer, this would annoy me.]

Table 4: Summarized volatility information of each cluster for Claims.

Volatility Level (Cluster)	Minimum	Mean	Maximum	σ
V1 (3)	51	79	154	**13**
V2 (4)	1,039	1,109	1,324	**52**
V3 (2)	221	1,687	8,841	**1,284**
V4 (1)	2	9,717	2,036,833	**64,835**

Coefficients of clustered results are used to calculate premiums in car insurance. Table 12 (Appendix B) shows the coefficients of the fitted model. In each cluster, statistical significance varies but overall the majority of coefficients are statistically significant.

In summary, the drivers have been clustered into four categories with distinct characteristics outlined in Table 4. We have seen how using the results from GCWM, one can create an insurance model based on clustering algorithms with various levels of risk represented in each cluster. GCWM found a group that contains a clear majority of drivers, in which the volatility of their claims was extremely low regardless of $Density$ or $DriverAge$. The results show that GCWM may potentially find unique clusters that are otherwise hidden within the data.

4.2.2 Modeling Claims Frequency

In this section, we model frequency of the French motor claims. We consider the covariates $Density$, $DriverAge$, $CarAge$, $Exposure$ and $Power$. The choice of covariates stems from the previously modelled single compo-

nent ZIP (Charpentier, 2014). The GCWM is fitted with the following expression

$$g_p(\mathbb{E}\left[Y_{ClaimNb}|\boldsymbol{x}\right]) = \beta_{j0} + \beta_{jDensity}x_{Density} + \beta_{jExposure}x_{Exposure} + \beta_{jPower}x_{Power} =: \tilde{\boldsymbol{x}}\boldsymbol{\beta}_j', \quad (4.2)$$

$$g_z(\mathbb{E}\left[Y_{ClaimNb}|\boldsymbol{x}\right]) = \bar{\beta}_{j0} + \bar{\beta}_{jExposure}x_{Exposure} + \bar{\beta}_{jCarAge}x_{CarAge} =: \tilde{\boldsymbol{x}}\bar{\boldsymbol{\beta}}_j', \quad (4.3)$$

[Same comment as before about the way this is presented.] where $Density$, $Exposure$, and $Power$ are explanitory variables in a Poisson model, while $Exposure$ and $CarAge$ are explantory for a Bernoulli model. As in Section 4.2.1, we also impose a log-normal assumption on the $Density$ covariate. The link functions g_p and g_z are chosen to be the log link and logit link respectively as in (2.6). After fitting the model, GCWM has found two zero-inflated components and one Poisson component as the best model to represent the data. The size of each cluster is displayed in Table 5 and we note a fairly even spread of the size across the three clusters.

Table 5: Size of clusters for the GCWM a model. [There is clearly something other than sizes in this table, i.e., the colours.]

Cluster 1	Cluster 2	Cluster 3
100492	163503	149174
Red	Green	Blue

Similarly to modelling severity, the GCWM finds clusters with unique characateristics. This is evident when looking at the Claims vs. Density plot (Figure 3). We see that the GCWM has split up the drivers into three groups based on the Density of cities. Table 6 shows that Cluster 2 drivers live in the least dense cities with a mean of 7.86 km on the log-scale. Followed by Clusters 3 and 1 with a mean of 5.23 km and 3.38 km, respectively.

Figure 3: Showing clusters in color for Frequency vs Density under log-normal assumptions.

Table 6: Summary of each cluster with log-normal assumptions for the $Density$ covariate measured in km on a log scale.

Cluster	Color	Minimum	Mean	Maximum	σ
1	Red	0.69	3.38	5.60	0.60
2	Green	6.29	7.86	10.20	1.03
3	Blue	4.13	5.23	9.66	0.65

Table 13 (Appendix B) shows a summary of the coefficients for the zero-inflated model. The significance codes are the same as in Table 12. In each cluster we can see that the majority of the coefficients are significant.

Specifically for the coefficients pertaining to the the Bernoulli zero-count models. Furthermore, by using the likelihood ratio test defined in (3.8), Cluster 1 has no evidence of zero-inflation. Thus, only a Poisson model was estimated for Cluster 1. [Reword previous sentence.] In summary, we see that the GCWM can account for zero-inflated data in pricing.

5   Simulation Study

Two simulation studies are conducted to determine the validity of the log-normal assumption and the effectiveness of the Bernoulli-Poisson partitioning method. The first section outlines the need for a non-Gaussian assumption for the covariates. The second section shows the classification accuracy and other relevant analysis for the Bernoulli-Poisson method.

5.1   Simulation Study for GCWM

In this section, we show how the proposed methodology works for different simulation settings. The simulation study is generated based on the regression coefficients of the CASdataset used in the previous section. The aim of the simulation study is to test the accuracy and ability of both GCWM a and CWM to return estimates of true parameters when one or more of the covariates is log-normal and the other two are Gaussian. This design specifically tests both models in the event when one of the covariates is non-Gaussian. The motivation behind this is fact is that many covariates used in insurance are likely to come from non-Gaussian distributions. Thus this simulation is aimed to test the relevancy of CWM, which treats all covariates as Gaussian. We define Model 1 as the baseline model, where the coefficients are selected to be reminiscent of those estimated from **CASdataset** and reported in upper portion of Table 12. The intercept for each component is increased to make sure simulated loss is positive. For ease of interpretation, these coefficients are then rounded and treated as true parameters for the simulation study. A simulation with a 3 group mixture model is generated from the aforementioned true parameters however, the third covariate (X_3) is generated from a log-normal distribution. In addition, the covariate X_2 for the second component is made insignificant and has no effect on the response. . [Some problems with the previous sentence: (1) I think we mean "Model 1 is the baseline model, where the coefficients. . .''; (2) I don't fully understand the reference to Table 2; and (3) Table 2, as with all other tables and figures, should be referenced in LaTeX via a nickname. I won't repeat the latter comment but please make sure all tables and figures are referenced using nicknames.] [Please read the last three sentences carefully.]

[Do not use past tense anywhere but the conclusions section. I have made some edits but I am sure I missed some others.] Results aggregated from 1000 runs are summarized in Table 7 and Table 8 . Given a

17

Gaussian assumption of for the residual error, we record the percentage of runs in which the error fall between a two-tailed %5 confidence interval in Table 7. For example, we report 90.10% accuracy for predictor X_2 in the component 2. This means that 90.1% of the time the true parameter is estimated within a 95% confidence interval. Further, we create Models 2, 3, 4 and 5 by altering the parameters of Model 1 by +30%, −30%, +50%, and −50%, respectively, and keeping the second covariate of the second component as an insignificant predictor form the CASdataset model. This is done to test the accuracy of the GCWM and its sensitivity to changes in coefficient sizes. Based on the results in Table 7, we can see that GCWM performs well for all simulation settings.

Table 7: GCWM a vs CWM Accuracy: covariate X_3 is treated as log-normally distributed, while the rest of covariates are of the Gaussian type.

Mod.	G	Int.	X_1	X_2	X_3	Int.	X_1	X_2	X_3
1	1	93.00%	90.10%	93.00%	93.10%	0.00%	0.00%	0.00%	0.00%
	2	90.10%	0.00%	90.10%	90.10%	0.00%	0.00%	0.00%	0.00%
	3	99.20%	99.10%	99.20%	99.20%	0.00%	0.00%	0.00%	0.00%
2	1	89.80%	89.20%	89.80%	89.80%	0.00%	0.00%	4.60%	0.00%
	2	89.20%	0.00%	89.20%	89.20%	0.00%	0.00%	0.00%	0.00%
	3	99.20%	99.20%	99.20%	99.20%	0.00%	0.20%	1.70%	0.00%
3	1	100.00%	100.00%	100.00%	100.00%	0.00%	0.00%	0.00%	0.00%
	2	100.00%	0.00%	100.00%	100.00%	0.00%	0.00%	0.00%	0.00%
	3	99.20%	99.20%	99.20%	99.20%	0.00%	0.00%	0.00%	0.00%
4	1	88.60%	86.80%	88.60%	87.00%	0.00%	0.00%	0.00%	0.00%
	2	86.90%	0.00%	86.90%	86.90%	0.00%	0.00%	0.00%	0.00%
	3	99.20%	99.20%	99.20%	99.20%	0.00%	0.00%	0.00%	0.00%
5	1	85.90%	84.90%	85.60%	85.90%	0.00%	0.00%	0.00%	0.00%
	2	85.00%	0.00%	84.90%	84.90%	0.00%	0.00%	0.00%	0.00%
	3	99.20%	99.20%	99.20%	99.20%	0.00%	0.20%	10.90%	0.00%

In line with expectations we note that barely any of the simulation runs are estimated correctlly, as most of the results are zero. This means that the performance of CWM approach is poor in presence of one non-Gaussian covariate which in this case is a log-normal covariate.

Table 8 provides the summary of mean squared errors (MSEs). The MSE is computed using the following formula, MSE $(\beta) = \frac{\sum_i^n (\beta_i - \hat{\beta}_i)^2}{n}$. Here, n accounts for the number of simulation runs, β is the true parameter of interest while $\hat{\beta}$ accounts for its estimate. The MSE for each parameter pertaining to each model are aggregated in Table 7. [Try to avoid footnotes.] Overall GCWM outperforms CWM in both accuracy and MSE.

We can observe that the MSEs for most of the models and their corresponding coefficients are not calculated at all due to convergence failures and as such they are shown as $(\cdot)$. This is not surprising because Table 7 shows

Table 8: GCWM results: the summary of MSE for all parameters used in five models. The covariate X_3 is treated as log-normal distributed, while the rest of covariates are Gaussian. These results correspond to same simulated runs as those in Table 7.

Mod.	G	β_o	MSE(β_o)	β_1	MSE(β_1)	β_2	MSE(β_2)	β_3	MSE(β_3)
1	1	1028	(11.353)	0.03	(0.00)	3.5	(0.00)	-380	(0.09)
	2	1600	(0.000)	-0.01	(0.00)	1.5	(0.00)	-250	(0.00)
	3	40000	(0.035)	-6.00	(0.00)	-305	(0.00)	1100	(0.47)
2	1	1350	(0.167)	0.04	(0.00)	4.5	(0.00)	-500	(0.03)
	2	2080	(0.001)	0.04	(0.00)	2.0	(0.00)	-325	(0.00)
	3	52000	(0.012)	-8.00	(0.00)	450	(0.00)	14300	(0.01)
3	1	720	(0.001)	0.02	(0.00)	2.5	(0.00)	-266	(0.00)
	2	1100	(0.008)	0.00	(0.00)	1.1	(0.00)	-17511	(0.00)
	3	28000	(0.002)	-4.20	(0.00)	245	(0.00)	7700.	(0.00)
4	1	1650	(13.056)	0.05	(0.00)	5.3	(0.00)	-570	(0.00)
	2	2400	(0.000)	-0.01	(0.00)	2.3	(0.00)	-375	(0.00)
	3	60000	(0.051)	-9.00	(0.00)	-457	(0.00)	16500	(0.00)
5	1	500	(1.115)	0.02	(0.00)	2.0	(0.00)	-190	(0.05)
	2	800	(0.003)	0.00	(0.00)	0.8	(0.00)	-120	(0.00)
	3	20000	(0.000)	-3.00	(0.00)	-150	(0.00)	5500	(0.00)

Table 9: CWM results: the summary of MSE for all parameters used in five models. All three covariates are treated as Gaussian. These results correspond to same simulated runs as those in Table 7.

Mod.	G	β_o	MSE(β_o)	β_1	MSE(β_1)	β_2	MSE(β_2)	β_3	MSE(β_3)
1	1	1028	(·)	0.03	(·)	3.5	(·)	-380	(·)
	2	1600	(·)	-0.01	(·)	1.5	(·)	-250	(·)
	3	40000	(·)	-6.00	(·)	-305	(·)	1100	(·)
2	1	1350	(·)	0.04	(·)	4.5	(·)	-500	(·)
	2	2080	(·)	0.04	(·)	2.0	(·)	-325	(·)
	3	52000	(·)	-8.00	(0.006)	450	(44.1)	14300	(·)
3	1	720	(·)	0.02	(·)	2.5	(·)	-266	(·)
	2	1100	(65.814)	0.00	(·)	1.1	(·)	-17511	(·)
	3	28000	(·)	-4.20	(·)	245	(·)	7700.	(·)
4	1	1650	(·)	0.05	(·)	5.3	(·)	-570	(·)
	2	2400	(·)	-0.01	(·)	2.3	(·)	-375	(·)
	3	60000	(·)	-9.00	(·)	-457	(·)	16500	(·)
5	1	500	(·)	0.02	(·)	2.0	(·)	-190	(·)
	2	800	(·)	0.00	(·)	0.8	(·)	-120	(·)
	3	20000	(·)	-3.00	(0.003)	-150	(4.7)	5500	(·)

the accuracy of CWM is not good when attempting to model non-Gaussian predictors as Gaussian. [Why are we talking about Tables 4, 5 and 6 here? Again, nicknames should be used.]

In summary, our simulation results showed good performance of the GCWM approach in modeling non-Gaussian covariates. More specifically, these results show high accuracy when covariates are log-normal. In contrary, CWM fails to estimate parameters accurately when the Gaussian assumption is violated.

5.2   Simulation Study - Bernoulli-Poisson Partitioning

In this section we show how the Bernoulli-Poisson (BP) partitioning method behaves under different conditions. The components are genereated under similar coefficients estimated from the **CASDatasets** package. Again for easy of interpretation, coefficients are rounded and treated as true parameters from which the simulated data is generated from. The mean and standard deviation of the covariates within each component was also taken into account when generating data. The first simulation examines the performance of the GCWM model for classification. We generate three components each with sample size $N = 1000$ for a total of 3000 simulated points. The model generated is similar to the mean and standard deviations of Table 6. Consider three simulated covariates with

$$g_p(\mathbb{E}\left[Y_{SimClaimNb}|\boldsymbol{X}\right]) = \beta_0 + \beta_{SimDensity}X_{SimDensity} + \beta_{SimDriverAge}X_{SimDriverAge}$$
$$+ \beta_{SimCarAge}X_{SimCarAge}, \tag{5.1}$$

$$g_z(\mathbb{E}\left[Y_{SimClaimNb}|\boldsymbol{X}\right]) = \bar{\beta}_0 + \bar{\beta}_{SimDensity}X_{SimDensity} + \bar{\beta}_{SimDriverAge}X_{SimDriverAge}$$
$$+ \bar{\beta}_{SimCarAge}X_{SimCarAge}. \tag{5.2}$$

as their respective linear models. The covariates $X_{SimDensity}$, $X_{SimDriverAge}$, and $X_{SimCarAge}$ are considered for both the Poisson and Bernoulli models. Furthermore the link functions g_p and g_z are chosen to be log-link and logit-link respectively. Here, the GCWM classifies drivers based on simulated data into three components. The misclassification rate is calculated by the proportion of true labels placed in other components by the GCWM a model. The results of the simulation are aggregated in Table 10. We observe the total misclassification rate of 1.8%, where the majority of misclassified components are between components two and three.

Table 10: Misclassfication rate and label comparison of generated data.

True Labels	Classified			Misclassification Rate
	1	2	3	
1	992	3	5	0.80 %
2	0	990	10	1.00 %
3	15	20	965	3.50 %
Overall Misclassification Rate				1.80 %
Average Purity				98.23 %
Adjusted Rand Index				0.9479

The simulation is expanded further to show how Bernoulli-Poisson partitioning behaves over 1000 runs and under two different conditions. The first condition is defined as follows. The mean and standard deviations of covariates are taken directly from sample statistics of **CASDataset**. The second condition involves adjusting the means of two of the covariates so they are closer to each other. The goal is to show that the BP-method holds its use even when means among covariates are close. The conditions are divided into two scenarios. In the first scenario which we consider "normal", the covariate means are taken directly from the sample data. In the second scenario defined as "close", the covariate means are manipulated so that they are closer to each other within some degree. This is a common problem in classification where if the means among two different components are close, then misclassification rate increases (Lim et al., 2014). The expansion of the simulation to 1000 runs tests the accuracy of 3 different partitioning methods to initialize a zero-inflated model. The results of this expansion are aggregated in Table 11. The Poisson partitioning method assumes that the presence of non-zeros will provide a better partitioning of the data-set. The Bernoulli partitioning method assumes that the presence of excess zeros will determine the best partitioning of the data-set. Finally the BP partitioning method assumes that both methods are weighed equally and therefore both must be taken into account when partitioning the dataset. The mean and standard deviation of each measurement is provided in Table 11.

Table 11: Aggregated results for the 1000 run simulation, mean and standard deviations for each statistic are compared across three methods.

Type	Condition	Poisson	(σ)	Bernoulli	(σ)	BP	(σ)
Misclassification Rate	normal	1.70%	(6.00)	1.60%	(6.00)	1.10%	(0.02)
	close	5.00%	(7.00)	6.00%	(2.00)	7.00%	(4.00)
Average Purity	normal	98.87%	(2.00)	98.91%	(2.25)	99.18%	(0.81)
	close	95.38%	(4.00)	94.55%	(1.00)	96.95%	(0.48)
Adjusted Rand Index	normal	0.9662	(0.07)	0.9677	(0.07)	0.9729	(0.0217)
	close	0.8706	(0.08)	0.8366	(0.04)	0.8538	(0.0453)

Under condition normal, the BP method shows better performance in error and is found to be less sensitive than other methods with an error rate of 1.10% and a standard deviation of 0.02%. Furthermore, when the close condition is imposed, the partitioning using only the Bernoulli method has better performance in terms of accuracy. The Adjusted Rand Index (ARI) is defined as the corrected for chance of the number of object pairs that are either in the same group or in different groups across all partitions divided by the total number of object pairs. The ARI measurements across all methods are promising. In particular the BP partitioning method under the normal condition has a very good ARI with a small standard deviation. The Average Purity (AP) is calculated as the average of the diagonal classification entries. The AP for the BP partitioning method

is the best out of all other methods, therefore the BP method is the most relevant for classification.

6 Conclusion

In this paper, we extend the class of generalized linear mixture CWM models by accomplishing two main goals. First, we proposed the methodology that allows for continuous covariates to follow a non-Gaussian distribution. Imposing Gaussian distribution on a skewed data may result in an suboptimal model fit. Second, we proposed a new Poisson CWM methodology that uses Bernoulli-Poisson partitioning method and allows for implementation of zero-inflated Poisson CWM model (ZI-GCWM). We name our proposed model class as the Generelized Cluster-Weighted Model (GCWM), to reflect the two extensions made to the existing CWM class of models.

Our proposed GCWM models allow for great applications in predictive modeling of insurance claims by overcoming a few limitations of the current CWM models. The ZI-GCWM allows for finding clusters within claims frequency which is an important information in risk classification and modeling of claims frequency. Further, some insurance rating variables used in the predictive modeling of severity claims may not strictly follow Gaussian assumptions, for example driver's age or car age, when treated as continuous covariates. An adequate extension to non-Gaussian covariates can be considered to relax current assumptions and improve the model fit. Given our data, we convincingly demonstrated that there is a need for a log-normal assumption in the $Density$ covariate, and by making it we have consideredly the model fit.

The results of our extensive simulation study showed the excellent performance of the proposed models in case of modeling non-Gaussian covariates. We found that current CWM model fails to estimate the parameters accurately when the Gaussian assumption is violated. The GCWM a shows significant improvement in the model fit over the CWM model based on AIC and BIC criteria. We also tested Bernoulli-Poisson partitioning of zero-inflated GCWM under different conditions and found that our proposed partitioning method has a very low misclassification rate, high average purity, and high average rand index.

Our approach is relevant to the actuarial pricing and risk management when current practices are based on implementation of various GLM models. Further extension of this work may incorporate modifications of the CWM family to allow for modeling limited depended variable or the right-censored data structure (refer to Miljkovic and Barabanov (2015) and Miljkovic and Orr (2017)).

A Derivation of the Log-normal Distribution

Consider a random variable U having univariate log-normal distribution with parameters $\mu \in \mathbb{R}$ and $\sigma \in \mathbb{R}_+$. Have $u \in \mathbb{R}_+$, then the probability density function of random variable U is defined as [2]

$$\mathcal{LN}(u; \mu, \sigma) = \frac{1}{u\sigma\sqrt{2\pi}} \exp\left[-\frac{(\ln u - \mu)^2}{2\sigma^2}\right].$$

Further, if random variable X is normally distributed i.e. $X \sim \mathcal{N}(x; \mu, \sigma)$, then $U := \exp(X) \sim \mathcal{LN}(u; \mu, \sigma)$. To see this, let $p_U(u)$, and $p_X(x)$ be the probability density functions of U and X respectively. By the change of variables theorem (see Murphy and Bach (2012) section 2.6.2.1) the density $p_U(u)$ is derived as

$$p_U(u) = p_X(\ln u)\frac{\partial}{\partial u}\ln u = p_X(\ln u)\frac{1}{u} = \frac{1}{u\sigma\sqrt{2\pi}} \exp\left[-\frac{(\ln u - \mu)^2}{2\sigma^2}\right].$$

We extend to a log-normal multivariate case where the random variable $\boldsymbol{U}$ is parameterized by $\boldsymbol{\mu} \in \mathbb{R}^p$ and $\boldsymbol{\Sigma} \in \mathbb{R}_+^{p \times p}$.

Lemma A.1. *Let the random variable $\boldsymbol{X}$ have multivariate normal distribution ie. $\boldsymbol{X} \sim \mathcal{MVN}(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $\boldsymbol{U} := \exp(\boldsymbol{X}) \sim f^U(\boldsymbol{u}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$. Here have $\boldsymbol{u} \in \mathbb{R}_+^p$ and the probability density function f^U is*

$$f^U(\boldsymbol{u}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(\prod_{i=1}^p u_i)|\boldsymbol{\Sigma}|(2\pi)^{\frac{p}{2}}} \exp\left[-\frac{1}{2}(\ln \boldsymbol{u} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\ln \boldsymbol{u} - \boldsymbol{\mu})\right].$$

Proof. Let $f^U(\boldsymbol{u}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $f^X(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ be the probability density functions of $\boldsymbol{U}$ and $\boldsymbol{X}$ respectively. By the multivariate change of variables theorem (see Murphy and Bach (2012) section 2.6.2.1), we derive the log-normal distribution, where $|\det J_{\ln}(u)|$ is the absolute value of the determinant for the Jacobian of the multivariate transformation $\ln(\boldsymbol{U}) = \boldsymbol{X}$. Hence,

$$|\det J_{\ln}(\boldsymbol{u})| = \prod_{i=1}^p u_i^{-1}, \text{ and}$$

$$\begin{aligned}
f^U(\boldsymbol{u}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) &= f^X(\ln \boldsymbol{u}; \boldsymbol{\mu}, \boldsymbol{\Sigma})|\det J_{\ln}(u)| \\
&= f^X(\ln \boldsymbol{u}; \boldsymbol{\mu}, \boldsymbol{\Sigma})\prod_{i=1}^p u_i^{-1} \\
&= \frac{1}{(\prod_{i=1}^p u_i)|\boldsymbol{\Sigma}|(2\pi)^{\frac{p}{2}}} \exp\left[-\frac{1}{2}(\ln \boldsymbol{u} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\ln \boldsymbol{u} - \boldsymbol{\mu})\right].
\end{aligned}$$

$\square$

---

[2]For full definition see Johnson et al. (1994)

Table 12: Summary of coefficients for severity clusters.

Coefficient [a]	V1 (Red) Estimate	Error	P	V2 (Green) Estimate	Error	P	V3 (Blue) Estimate	Error	P	V4 (Teal) Estimate	Error	P
Intercept	7.876	0.137	***	7.180	0.061	***	4.673	0.014	***	7.077	0.003	***
Density	-0.031	0.009	***	0.005	0.004		-0.011	0.001	***	0.002	0.002	
C2	-0.172	0.080	*	0.064	0.034	.	0.020	0.001	**	0.008	0.002	***
C3	-0.396	0.080	***	0.108	0.034	**	0.010	0.007		0.003	0.002	.
C4	-0.642	0.081	***	-0.033	0.035		0.034	0.007	***	0.005	0.002	**
C5	-0.500	0.090	***	0.066	0.039	.	0.069	0.007	***	0.011	0.002	***
D2	-0.535	0.083	***	-0.168	0.038	***	-0.217	0.009	***	-0.006	0.001	***
D3	-0.607	0.084	***	-0.241	0.038	***	-0.205	0.009	***	-0.008	0.001	***
D4	-0.390	0.099	***	-0.122	0.045	**	-0.200	0.0106	***	-0.009	0.002	***
D5	0.123	0.101		0.035	0.047		-0.138	0.010	***	-0.002	0.002	
R23	0.003	0.131		-0.016	0.053		-0.001	0.012		0.002	0.006	
R24	-0.232	0.054	***	-0.017	0.023		-0.102	0.005	***	-0.015	0.013	***
R25	0.144	0.096		-0.184	0.043	***	-0.065	0.009	***	-0.016	0.024	***
R31	-0.009	0.073		0.055	0.031	.	-0.141	0.008	***	-0.003	0.018	.
R52	-0.303	0.064	***	0.012	0.028		-0.142	0.006	***	-0.015	0.038	***
F53	-0.153	0.063	*	0.095	0.028	***	-0.012	0.006	*	-0.014	0.001	***
R54	-0.222	0.082	**	0.074	0.037	*	-0.122	0.007	***	-0.015	0.002	***
R72	-0.098	0.072		0.175	0.031	***	-0.081	0.007	***	-0.007	0.002	***
R74	-0.236	0.142	.	-0.114	0.067	.	0.466	0.016	***	-0.019	0.003	***
P-FGH	0.123	0.033	***	0.012	0.015		0.001	0.003		0.002	0.001	*
P-Other	0.131	0.045	**	0.075	0.020	***	0.012	0.003	**	0.005	0.001	***
GR	-0.095	0.031	**	-0.029	0.014	*	0.005	0.002	.	-0.005	0.001	***

[a] The significance codes are defined as $P < 0.001$: (***), $0.001 < P < 0.01$: (**), $0.01 < P < 0.05$: (*), $0.05 < P < 0.10$: (.) pertaining to the P value of the specific coefficient. C# refers to the Car Age category, D# refers to the Driver Age category, R# refers to the region of France, and P refers to the power category.

Table 13: Summary of coefficients for frequency clusters.

Coefficient [a]	Cluster 1 (Red) Estimate	Error	P	Cluster 2 (Green) Estimate	Error	P	Cluster 3 (Blue) Estimate	Error	P
(Intercept)	-10.199	0.109	***	-7.217	0.216	***	-13.694	0.099	***
Density	1.860	0.027	***	0.442	0.013	***	1.771	0.013	***
Exposure	0.825	0.040	***	0.726	0.192	***	0.745	0.046	***
P-GH	-0.063	0.030	*	-0.004	0.035		0.013	0.030	
P-Other	-0.019	0.042		0.049	0.043		0.107	0.040	**
(Intercept)				2.424	0.188	***	1.886	0.356	***
Exposure				-5.871	0.758	***	16.520	2.990	***
C2				-0.541	0.176	**	-0.922	0.368	*
C3				-1.217	0.201	***	-2.263	0.553	***
C4				-1.265	0.225	***	10.805	72.677	
C5				-0.808	0.266	**	-6.355	30.867	

[a]The significance codes are defined as $P < 0.001$: (***), $0.001 < P < 0.01$: (**), $0.01 < P < 0.05$: (*), $0.05 < P < 0.10$: (.) pertaining to the P value of the specific coefficient.

References

Bermúdez, L., Karlis, D., 2012. A finite mixture of bivariate poisson regression models with an application to insurance ratemaking. Computational Statistics and Data Analysis 56 (12), 3988–3999.

Biernacki, C., Celeux, G., Govaert, G., Jul 2000. Assessing a mixture model for clustering with the integrated completed likelihood. IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (7), 719–725.

Brown, G. O., Buckley, W. S., 2015. Experience rating with poisson mixtures. Annals of Actuarial Science 9 (02), 304–321.

Charpentier, A., 2014. Computational Actuarial Science with R. CRC press.

Dempster, A. P., Laird, N. M., Rubin, D. B., 1977. Maximum likelihood from incomplete data via the EM-algorithm. Journal of the Royal Statistical Society B 39, 1–38.

Dutang, C., Charpentier, A., 2016. CASdatasets. R package version 1.0-6.
URL http://cas.uqam.ca/pub/R/

Gershenfeld, N., 1997. Nonlinear inference and cluster-weighted modeling. Annals of the New York Academy of Sciences 808 (1), 18–24.

Gershenfeld, N., Schoner, B., Metois, E., 1999. Cluster-weighted modelling for time-series analysis. Nature 397 (67171), 329–332.

Gershenfeld, N. A., 1999. The nature of mathematical modeling. Cambridge university press.

Ingrassia, S., Minotti, S. C., Punzo, A., 2014a. Model-based clustering via linear cluster-weighted models. Computational Statistics & Data Analysis 71, 159–182.

Ingrassia, S., Minotti, S. C., Vittadini, G., 2014b. Local statistical modeling via a cluster-weighted approach with elliptical distributions. Journal of classification 29 (3), 363–401.

Ingrassia, S., Punzo, A., Vittadini, G., Minotti, S. C., Jul 2015. Erratum to: The generalized linear mixed cluster-weighted model. Journal of Classification 32 (2), 327–355.

Johnson, N. L., Kotz, S., Balakrishnan, N., 1994. Continuous Univariate Probability Distributions,(Vol. 1). John Wiley & Sons Inc., NY.

Lambert, D., 02 1992. Zero-inflated poisson regression, with an application to defects in manufacturing. Technometrics 34, 1–14.

Lee, S. C. K., Lin, X. S., 2010. Modeling and evaluating insurance losses via mixtures of Erlang distributions. North American Actuarial Journal 14 (1), 107–130.

Lim, H., Li, W., Yu, P., 03 2014. Zero-inflated poisson regression mixture model. Computational Statistics and Data Analysis 71, 151?158.

McCullagh, P., Nelder, J. A., 1989. Generalized linear models. Vol. 37. CRC press.

McNicholas, P. D., 2016. Mixture Model-Based Classification. Chapman & Hall/CRC Press, Boca Raton.

McNicholas, P. D., Murphy, T. B., Sep 2008. Parsimonious gaussian mixture models. Statistics and Computing 18 (3), 285–296.
URL https://doi.org/10.1007/s11222-008-9056-0

Miljkovic, T., Barabanov, N., 2015. Modeling veterans health benefit grants using the expectation maximization algorithm. Journal of Applied Statistics 42 (6), 1166–1182.

Miljkovic, T., Fernández, D., May 2018. On two mixture-based clustering approaches used in modeling an insurance portfolio. Risks 6 (2), 57.

Miljkovic, T., Grün, B., 2016. Modeling loss data using mixtures of distributions. Insurance: Mathematics and Economics.

Miljkovic, T., Orr, M., 2017. An evaluation of the reconstructed coefficient of determination and potential adjustments. Communications in Statistics-Simulation and Computation 46 (9), 6705–6718.

Murphy, K. P., Bach, F., 2012. Machine Learning: A Probabilistic Perspective. Adaptive Computation and Machi. MIT Press.

Punzo, A., Ingrassia, S., ., 2014. Parsimonious generalized linear Gaussian cluster-weighted models. Springer International Publishing.

Punzo, A., McNicholas, P. D., 2017. Robust clustering in regression analysis via the contaminated gaussian cluster-weighted model. Journal of Classification 34 (2), 249–293.

R Core Team, 2018. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Subedi, S., Punzo, A., Ingrassia, S., McNicholas, P. D., 2013. Clustering and classification via cluster-weighted factor analyzers. Advances in Data Analysis and Classification 7 (1), 5–40.

Subedi, S., Punzo, A., Ingrassia, S., McNicholas, P. D., 2015. Cluster-weighted t-factor analyzers for robust model-based clustering and dimension reduction. Statistical Methods and Applications 24 (4), 623–649.

Verbelen, R., Gong, L., Antonio, K., Badescu, A., Lin, S., 2015. Fitting mixtures of Erlangs to censored and truncated data using the EM algorithm. ASTIN Bulletin 45 (3), 729–758.

Vuong, Q. H., 1989. Likelihood ratio tests for model selection and non-nested hypotheses. Econometrica 57 (2), 307–333.

Wedel, M., 2002. Concominat variables in finite mixture modeling. Statistica Neerlandica 56 (3), 362–375.

Wedel, M., De Sabro, W., 1995. A mixture likelihood approach for generalized linear models. Journal of Classification 12 (3), 21–55.

Wilson, P., 02 2015. The misuse of the vuong test for non-nested models to test for zero-inflation. Economics Letters 127.

Wilson, P., Einbeck, J., 2018. A new and intuitive test for zero modification. Statistical Modelling.